ChinaXiv合作期刊

基于电子商务评论的商家信誉维度构建*

王　宇　李秀秀
(大连理工大学管理与经济学部　大连　116024)

摘要: 【**目的**】通过对电子商务评论文本的分析和处理, 获取有效的商家信誉信息, 从客观角度建立商家信誉维度体系。【**方法**】基于 HNC 理论的同行优先原理和文本挖掘方法提出改进的评论文本主题词抽取方法和主题词聚类算法, 并进行类簇标签抽取及各类簇权重计算。【**结果**】生成商家信誉维度体系及各维度权重, 以京东平台手机评论文本为实例, 构建商家信誉维度体系, 并对其进行评价, 证明方法的可行性与有效性。【**局限**】受 HNC 词库不全的影响需手工生成一部分字词符号, 在应用到更大规模的评论文本处理时可能会存在限制。【**结论**】利用本文提出的方法建立的商家信誉维度体系能够客观地反映出用户真正关心的商品指标。

关键词: 评论文本　主题词聚类　信誉维度　电子商务
分类号: TP391

chinaXiv:201712.01380v1

1　引　言

近年来, 电子商务以及社交媒体蓬勃发展。最新数据显示, 2016 第三季度中国电子商务市场交易规模达到 5.2 万亿元, 同比增长 30.8%, 其中网络购物市场交易规模 1.15 万亿元, 同比增长 23.6%[1]; 2016 年社交媒体用户达到 23.1 亿人, 相当于全球人口的 31%, 新增社交媒体用户 2.19 亿人, 年增幅 10%[2]。

随着网上商家数量的快速增长, 商品种类、数量的极大丰富, 商家信誉状况却良莠不齐, 并有大量假货充斥其中, 加之商品评价信息多以非结构化的形式存在于网络中, 消费者很难仅从商家对商品的描述中辨别真伪, 做出正确购买决策。因此如何对评论短文本进行有效的分析和处理, 以获取有效的商家信誉信息, 从而建立商家信誉维度体系, 已经成为研究的热点问题。鲁文[3]基于相关理论模型从 4 个维度构建了包含 17 个量化指标的电子商务在线信誉的影响模型。茹永梅[4]运用层次分析法和模糊综合评价法对 O2O 电子商务中的商家信誉进行度量, 建立基于模糊理论的 O2O 电子商务商家信誉评估模型。吴维芳等[5]利用 Word2Vec 对酒店评论进行特征抽取和降维, 结合情感分析技术, 研究影响酒店用户满意度的因素。

但目前商家信誉评价研究大部分都专注于数值化的研究方式, 却忽视了客户的定性评论对卖者信誉度的影响。调查结果表明, 在电子商务交易决策过程中, 交易双方越来越重视社会网络中其他参与者(如朋友、其他消费者、意见领袖、第三方平台等)的评价, 原因在于这些评价能为商家改善服务、提高信誉水平提供参考, 为消费者做出购买决策提供依据。虽然赵学锋等[6-7]通过文本聚类对在线零售商的客户评论进行维度分析, 扩展原有的信誉维度。但时至今日, 电子商务迅速发展, 尤其在与社交网络互相融合之后, 使得评论文本越来越带有社会化的特征, 文本量巨大, 语言灵活随意, 文本长短不一, 且包含较多无关信息。这些特征使得简单的聚类方法在面对如此大规模的评论文本时聚类的效果和准确度都将大大降低。

针对现有的商家信誉评价指标体系的不足, 本文从用户评论的角度, 基于 HNC 理论[8], 利用 HNC 同行优先原则对大量用户评论文本抽取主题词, 将主题词映射到 HNC 字词库, 采用基于 HNC 的词语相似度计算

通讯作者: 王宇, ORCID: 0000-0001-9759-3313, E-mail: ywang@dlut.edu.cn。

*本文系国家自然科学基金重点项目“社会化商务中参与者的信誉与信任机理及交易决策研究”(项目编号: 71431002)的研究成果之一。

ChinaXiv合作期刊

改进传统的 CURE 算法，提出一种新的针对评论文本的主题词聚类方法；在此基础上构建商家信誉指标体系，并对这种方法构建的指标体系进行检验和评价。

2 评论文本主题词抽取

主题词即能够表达文本主题的规范化词语或词组。传统的主题词抽取方法主要针对长文本，但评论文本长度短，不存在标题、首末句等词语位置信息，并且句型不规范，往往隐藏主语。本文提出一种针对评论文本的主题词抽取方法。

2.1 主题词扩展

针对评论文本的特点[9]，依据词性、词频率和词共现对评论中的高频主题词进行初步抽取。考虑主题词的广泛性以及同义词合并中不可避免的不完善情况，抽取主题词不能完全排除低频词，需要主题词间的词频有一定的差异[10-11]。因此，对于已经初步抽取出的高频主题词，通过依存句法提取出修饰这些主题词的形容词，并按照词频排序，只保留高频形容词，再针对未提取出高频主题词的评论文本，提取该形容词修饰的名词作为主题词。例如评论文本“鞋子收到了，保暖性很好，鞋底很厚，超出预期。”名词集合为{鞋子，保暖性，鞋底}，“鞋子”是通用词将被删除，而“鞋底”、“保暖性”无法达到主题词初步抽取的词频要求，针对该评论文本，通过初步抽取无法抽取出主题词，则进入扩展主题词。假如抽取评论文本集合的高频形容词集为{满意，快，好，合适…}，抽取该条评论文本的形容词集为{好，厚}。可发现“好”包含在高频形容词集中，“厚”不包含在内。根据依存句法发现“好”修饰的名词为“保暖性”，则“保暖性”进入主题词集合，如图 1 所示。

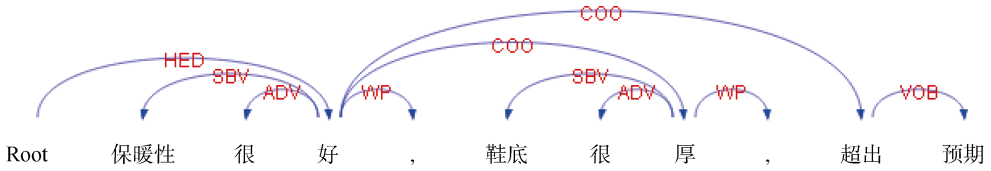

图 1 例句句法分析

2.2 词频调整

经过主题词扩展，可认为已经抽取出覆盖面足够广泛的主题词。对于仍然没有抽取出主题词的文本，有两种可能：一是评论文本确实不包含主题词，或词汇过于生僻，对于此种情况不作处理；二是由于评论文本句型不规范，隐藏了主题词或主语，对于这种情况利用 HNC“同行优先”原理进行处理。“同行优先”是 HNC 理论处理语义块内部语义距离的重要原则，可以简单理解为能够相互搭配或者相互修饰的词语具有相似的义项符号[12]。比如，“无私 uc3ae02”、“远大 gub01”、“奉献 vc3ae02”、“目标 grb01”，可以看出，“无私”和“奉献”以及“远大”和“目标”这两对常用搭配各自的 HNC 符号比较接近，与不搭配的词语义项符号则相差较远[13]。

利用这一特性，对于隐藏主语的评论文本，可提取该评论文本的形容词，比较该形容词与前面抽取出的主题词 HNC 符号相似度。当该形容词与某一主题词的 HNC 义项相似度超过设定阈值，并且与其他主题词同该形容词的相似度之差大于设定阈值，则认为该评论文本隐藏的主语为该主题词，该主题词词频加 1。

设抽取出的主题词集合为 $W=\{w_1, w_2, \cdots, w_n\}$，某条未抽取出主题词的评论文本包含形容词 a，则对于所有的 w_i，按照文献[14]提供的方法计算 HNC 相似度 $sim(a, w_i)$，并取最大值。若主题词 w_p 与形容词 a 的相似度最大，即 $sim(a, w_p)=max\{sim(a, w_i)\}$，且 $sim(a, w_p)>\alpha$，$sim(a, w_p)-sim(a, w_i)>\beta$，则主题词 w_p 的词频 $tf(w_p)=tf(w_p)+1$。例如，评论文本“挺满意的，很便宜，值得购买”经过主题词初步抽取以及主题词扩展，都不能抽取出主题词，则抽取该评论文本形容词集合为{满意，便宜}，假设抽取出的主题词集合为{质量，服务，物流，款式，价格…}，分别计算“满意”、“便宜”同主题词集合中各词汇的 HNC 相似度，发现“满意”同多个主题词的相似度在 0.4-0.5 之间，彼此差值很小，因此不能确定搭配主题词；“便宜”仅与主题词“价格”的相似度超过 0.9，同其他主题词的相似度均在 0.4 以下，因此认为“便宜”隐藏的主题词为“价格”，“价格”的词频加 1。

chinaXiv:201712.01380v1

ChinaXiv合作期刊

chinaXiv:201712.01380v1

3　评论文本主题词聚类

3.1　HNC 符号映射及主题词表示

HNC 理论以语义表达为基础，是一套完整、强大的语义网络描述体系。作为服务于汉语理解的语言知识库的重要组成部分，词知识库的建设也一直是 HNC 理论研究的重要工作。但目前包含 HNC 在内的各种词库如 WordNet、同义词词林等，都存在词库覆盖不全的问题。HNC 理论将概念(词汇)分为抽象概念和具体概念，抽象概念用五元组和语义网络表达，具体概念采用向抽象概念的基元概念和基本概念挂靠的方式表达。评论文本中抽取的词汇基本属于具体概念，对于这些抽取词汇中不包含在现有 HNC 词库中的部分，采用上述“类别符号+挂靠”的方式进行补充是可行的。

为了满足后续的聚类要求，设计一种基于 HNC 符号的主题词表示方法，即四元组表示法：{主题词，词频，HNC 符号，来源}。其中，主题词即主题词本身；词频是主题词在评论文本集中出现的总次数；HNC 符号是主题词映射到 HNC 字词库的 HNC 义项符号；来源是标识哪些评论文本包含该主题词或隐含该主题词或者经过同义词合并的近义词。

这样的表示方法为主题词引入了准确的语义信息，后续聚类过程的聚类对象就不再是简单的词形，而是含有语义的 HNC 符号，使得聚类结果更加精确。另外，保留主题词来源这一属性，可以在主题词聚类完成后，将主题词聚类簇还原为对应的评论文本类簇，便于对聚类簇的分析和描述。

3.2　主题词聚类算法

文本聚类有很多算法，比如划分法、层次法、密度法等，但适于对评论短文本聚类的算法却很少。CURE(Clustering Using REpresentatives)算法[15]采用多个代表点表示整个类簇，获得的类质量较高，并且在处理大数据量时采用随机取样、分区的方法提高其效率，比较适合用于评论文本的挖掘。但当簇的密度、分布不均匀时，CURE 算法会导致选取到不合理的代表点，造成不合理的簇合并。考虑到评论文本主题词的特性，本文提出基于HNC符号的改进CURE聚类算法。

(1) 代表点的选取算法

代表点影响因子和簇中心点的定义如下。

①代表点影响因子

设数据集簇 $C=\{d_1,d_2,\cdots,d_n\}$，其中 d_i 为簇中的数据点，n 为簇 C 中数据点个数，簇 C 的代表点集合为 $S(C)=\{d_{p1},d_{p2},\cdots,d_{pm}\}$，$m$ 为代表点个数，则代表点 d_{pj} 的影响因子为 $FIW(d_{pj})=\frac{|c_j|}{|c|}\times\frac{f_j}{N}$，其中 $|c_j|$ 为簇 C 中与代表点 d_{pj} 相似度最大的数据点个数，$|c|$ 为簇 C 中的主题词总数，f_j 为代表点主题词 d_{pj} 的词频，N 为类簇 C 中所有主题词的词频和，即 $N=\sum_{j=1}^{n}f_j$。

②簇中心点

设数据集簇 $C=\{d_1, d_2, \cdots, d_n\}$，其中 d_i 为簇中的数据点，n 为簇 C 中数据点个数，簇中心点 d_{mean} 是与其他主题词相似度的均值最大的点，即 d_{mean} 满足 $sim(d_{mean},d_{pj})=\max_{d\in C-\{d_{pj}\}}\left(\sum sim(d,d_{pj})/n\right)$，其中 $d_{mean}\in\{d\}$。

其中，簇中心点中代表点的相似度计算，采用文献[14]提出的基于 HNC 语义的相似度计算方法。

代表点的选取算法如下：

输入：数据集簇 $C=\{d_1,d_2,\cdots,d_n\}$，最大代表点个数 m，影响因子 FIW 阈值 η

输出：簇 C 的代表点集合 S_c

```
Begin
calculate d_mean(C)//计算簇中心点
initiate S_c={d_mean (C)}//选取簇中心点作为第一个代表点
    for each d_i in C- S_c {
        for each d_j in S_c {
            calculate sim_didj
            similarity.add(sim_didj)
        }
        //将 sim_didj 保存在临时数组 similarity 中
    }
    calculate max(similarity)//对于 C-S_c 中每一个数据点，计算
      其与已选代表点相似度的最大值
    for each sim_didj in similarity{
          if(sim_didj == max(similarity))
          max-similarity.add(sim_didj)
    }
    calculate min(max-similarity)//选取与已选代表点最大相似
      度值最小的点
    for each sim_didj in max-similarity
      if(sim_didj == min(max-similarity) and fiw(d_i)>η and
        length(S_c)<m)
        S_c = S_c ∪ {d_i}
End
```

其中，代表点的数量 m 根据原始数据集的数据量决定，在实际实验中影响因子 FIW 阈值 η 为 $\frac{1}{4m}$。

ChinaXiv合作期刊

(2) 基于 HNC 符号的改进 CURE 聚类算法

代表点选取规则改进后，在最终聚类簇的数量上，CURE 算法需要提前设定最终聚类簇的个数。改进算法不设置最终类簇的数目，而是通过控制簇合并时的相似度阈值 w 来调节类簇的合并，具体步骤如下:

①所有代表主题词的 HNC 符号 $\{h_1,h_2,\cdots,h_n\}$，对于每一个 h_i 创建一个簇 C_i，即 $C=\{C_1,C_2,\cdots,C_n\}$，$C_i=\{h_i\}$，C_i 的代表点集合 $S(C_i)=\{h_i\}$。

②如果簇集 C 的数目 $|C|<2$，执行终止。

③找出簇集 C 中距离最近的两个簇 C_u 和 C_v，如果 $dist(C_u,C_v)>w$，执行中止。

④合并簇 C_u、C_v，$C_{new}=C_u\cup C_v$，计算簇 C_{new} 的中心点，按上一节方法计算簇 C_{new} 的代表点集合 $S(C_{new})$。

⑤更新簇集 C: $C=C-C_u-C_v+C_{new}$，执行步骤②。

4 信誉维度体系构建

评论主题词经过聚类算法处理后得到的聚类簇集隐含着消费者关注的关于商家的信誉维度信息，对这些簇集进行标签抽取及命名即得到商家的信誉维度。另外，作为一个完整的维度体系，还需要为每个维度指标确定权重。

类簇集标签的抽取即从类簇中选择若干个具有代表性的词语表达整个类簇的主题。目前大多数类簇标签抽取方法都是简单地选取词频最高或者是 TF-IDF 值最大的若干个词语作为类簇标签[16-18]，但构成类簇标签的词语之间往往存在一定的关联性，而在上述关于类簇标签抽取的研究中，并没有考虑词语之间的关联性和逻辑关系。HNC 理论在构建词语的 HNC 符号时，通常是基于局部联想脉络，将概念之间存在关联的词语映射到相同或者相近的概念基元符号上，计算机通过解释相应的符号就可以把握概念之间的关联性。

基于 HNC 的“同行优先”原则考虑词语之间的关联概念节点(即概念基元)是否相同，将类簇中的词语划分为不同的词语集合，计算所有词语集合的权重，将权重最大的词语集合作为每个类簇的标签。同时，每个类簇标签所对应的词语集合的权重也就是相应的信誉维度权重。

某一个维度的权重体现了该维度在整个评价体系中的相对重要程度。由主题词簇集形成的指标体系，考虑每个聚类簇中主题词的数量 S、词语的词频 TF 以及词语与类簇中其他词语的语义相似度 $Sim(w_i,w_j)$ 总和的均值等三个因素综合评价维度的权重，其计算如公式(1)所示。

$$W_{weight}=\alpha_1\times S+\alpha_2\times TF+\alpha_3\times\overline{Sim}(w_i,w_j) \quad (1)$$

其中，$\overline{Sim}(w_i,w_j)=\dfrac{\sum_{i=1,i\neq j}^{C_{length}}Sim(w_i,w_j)}{C_{length}-1}$，$C_{length}$ 表示每个类簇中包含的词语个数。$\alpha_1+\alpha_2+\alpha_3=1$，根据实际经验和多次实验调整，这里 $\alpha_1,\alpha_2,\alpha_3$ 依次取 0.5, 0.3, 0.2。

如果某个维度中包含的主题词数量越多，包含的主题词频次越高，则代表它在更多的评论文本中被提及，被更多的消费者重视，权重应该越高，也就是说某一维度的重要性与其包含的主题词总量成正比。类簇标签抽取算法实现如下:

输入: 所有的类簇集合。

输出: 类簇标签集合。

①对类簇中的所有词语进行权重计算，对每个类簇任意选择一个词语 w_i 作为初始的词语集合。

②对每个类簇中剩余的词语逐个进行判断，首先判断是否和词语 w_i 拥有相同的关联概念节点，如果两个词语存在相同的关联概念节点，则将这个词语加入该词语集合中；如果不存在相同的关联概念节点，则依据 HNC 给出的概念关联式判断该词语与词语 w_i 是否存在某种逻辑关系，如果两个词语存在某种逻辑关系，则将这个词语加入该词语集合中。如果上述两种情况都不满足，则将该词语加入新的词语集合。

③对每个类簇中剩余的词语重复执行步骤②，直至所有词语都加入到相应的词语集合中。

④对所有生成的词语集合进行权重计算，选取权重最大的词语集合作为每个类簇的标签。

5 实验验证

5.1 实验设置

测试数据随机抓取于京东网站，共 1 850 条手机评论，筛选过滤字数过少以及无效的评论文本，剩余 1 000 条用于实验，由人工审阅提取主题词，对于省略主题词的文本，人为分配主题词并对提取出的主题词进行分类处理。其中主题词数量由于涉及到词频调整，因此从不同评论文本中抽取出的相同主题词，主题词

chinaXiv:201712.01380v1

ChinaXiv合作期刊

数量做累加。

实验主要分为两个部分，分别测试主题词抽取方法以及主题词聚类方法的效果。评价方法按照主题词抽取结果以及聚类结果与人工判断结果越吻合越好的原则，采用准确率、召回率对主题词抽取结果和聚类结果进行评估，定义如下。

(1) 主题词抽取：设人工审阅评论文本得到的主题词数量为 n，主题词抽取方法得到的主题词数量为 m，m 个主题词中与人工审阅结果吻合的主题个数为 e，则主题词抽取的准确率 P、召回率 R 的计算分别如公式(2)和公式(3)所示。

$$P=\frac{e}{n} \tag{2}$$

$$R=\frac{e}{m} \tag{3}$$

(2) 主题词聚类：设 $l(c_i)$是类簇 c_i 的簇标签，$l(d_j)$是第 j 个主题词人工标记的类别，n_i 是自动聚类簇 c_i 包含的主题词数目，m_i 是人工分类 d_j 包含的主题词数目，k 是类簇数目。主题词聚类的准确率 P’、召回率 R’的计算如公式(4)和公式(5)所示。

$$P'=\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\frac{1}{n_i}\delta(l(c_i),l(d_j)) \tag{4}$$

$$R'=\frac{1}{k}\sum_{i=1}^{k}\sum_{j=1}^{m_i}\frac{1}{m_i}\delta(l(c_i),l(d_j)) \tag{5}$$

其中，$\delta(x,y)=\begin{cases}0 & x\neq y\\ 1 & x=y\end{cases}$。

5.2　主题词抽取效果分析

实验中，分别使用传统基于词性的主题词抽取方法(初步抽取)、初步抽取+主题词扩展、 初步抽取+主题词扩展+基于 HNC 的词频调整三种方法，进行主题词抽取，检验三种方法的准确率、召回率，并对比分析。实验结果如图 2 和图 3 所示。

从上述主题词抽取的实验效果看，基于句法分析的主题词扩展以及基于 HNC 的词频调整，相对于初步抽取的结果，在准确率、召回率上都具有明显的提升。另外，三种抽取方法初期都显现出准确率与召回率随文本数量的增加而增加的特性，随后在一定范围内波动，加入基于 HNC 的词频整调的方法，表现出的增长性更加稳定，这主要是由于 HNC 对于隐藏主语的提取是在与前两步抽取出的主题词比对的基础上进

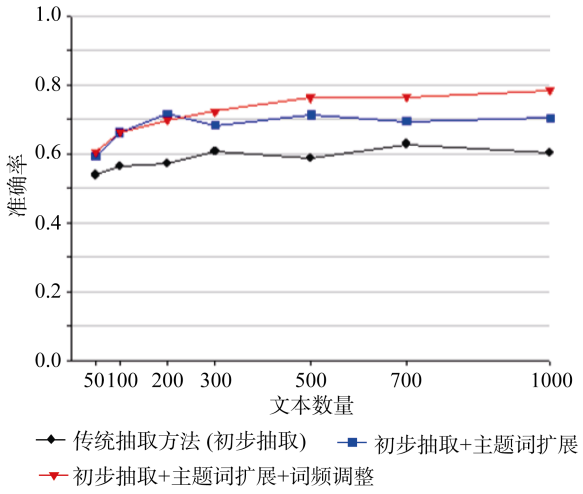

图 2　三种主题词抽取方法的准确率对比

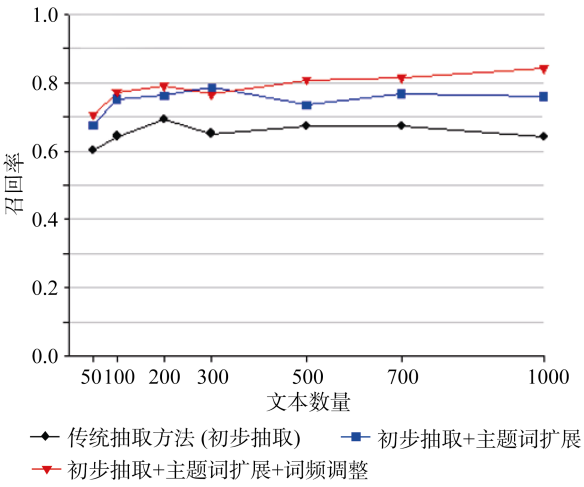

图 3　三种主题词抽取方法召回率对比

行的，文本量越多，抽取出的主题词越丰富，比对的效果也就越好。当然在不累计词频的情况下，后两种方法的效果是一致的。

5.3　聚类效果分析

实验中，分别使用传统的 CURE 方法+Jaccard 相似度计算方法、改进的 CURE 方法+基于知网的相似度计算方法、改进的 CURE 方法+基于 HNC 的相似度计算方法三种方法进行聚类，验证三种方法的准确率、召回率，并对比分析。实验结果如图 4 和图 5 所示。

从聚类实验效果看，本文提出的聚类算法在聚类的准确率和召回率上相对于传统 CURE 算法以及基于知网的算法都有提升，体现了算法改进的合理性以及 HNC 在语义相似度计算上的优势。另外，在时间复杂度上，改进后的 CURE 算法与传统 CURE 算

chinaXiv:201712.01380v1

ChinaXiv合作期刊

法一致，仍是 O(n2)，聚类准确率的提升并没有以时间为代价，对孤立点的处理也保持了传统 CURE 算法的优势。因此该方法比较适用于大规模的评论文本主题词聚类。

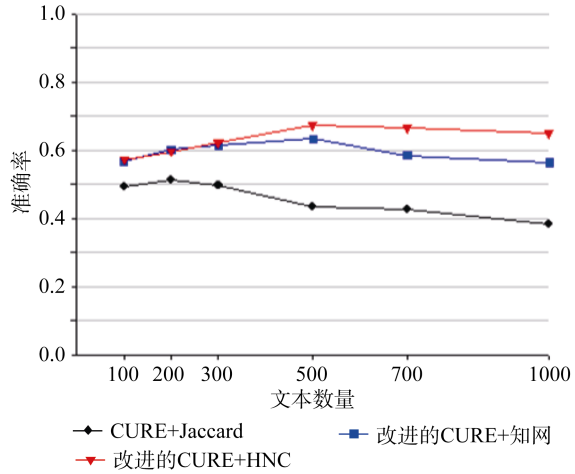

图 4　三种聚类方法准确率对比

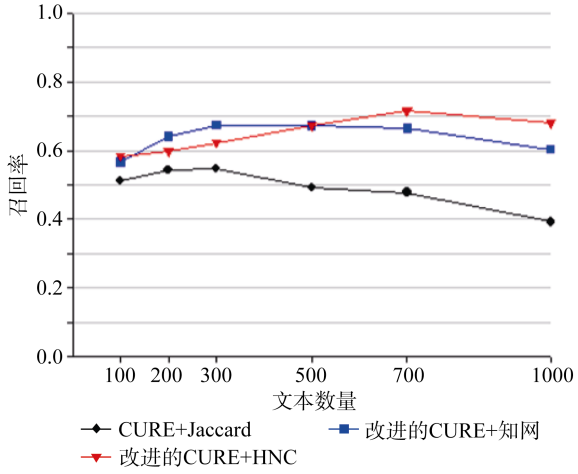

图 5　三种聚类方法召回率对比

6　实例分析

为了在实践中应用本文提出的商家信誉维度体系构建方法，以京东为研究平台，利用其提供的开放 API 抓取手机产品的评论语料 7 856 条，经过无效评论过滤筛选等处理，剩余有效评论语料 5 394 条。手机产品评论均来自京东自营，涉及 iPhone6、华为 P7、小米 2 等 15 种手机型号，评论时间跨度为 2014 年 5 月–2015 年 3 月，检索词为“手机”，筛选字段为“京东自营”，部分评论数据如图 6 所示。

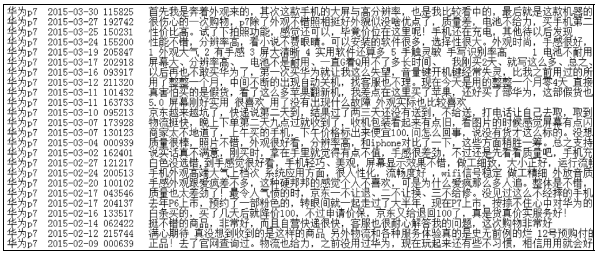

图 6　手机评论数据(部分)

(1) 主题词抽取。采用中国科学院计算技术研究所研发的 ICTCLAS2015 分词系统对评论文本进行分词并标注词性，使用第 2 节提出的主题词抽取方法，共抽取主题词 776 个(累计词频 6 187)，将抽取出的主题词与 HNC 字词库映射，以{主题词，词频，HNC 符号，来源}四元组形式存储。

(2) 主题词聚类。采用主题词聚类算法对评论文本中抽取出的主题词进行聚类，最终得到 9 个大类簇，35 个小类簇(词频累计数量小于 20)，另有 194 个词类别不确定或属于孤立点。

(3) 商家信誉维度体系构建。依据第 4 节方法为聚类得到的大类簇进行标签抽取，确定前 6 个类簇描述作为评价指标，以此建立信誉指标体系，并计算 6 个维度的权重，其结果如表 1 所示。

表 1　聚类结果

序号	维度名称	簇标签	主题词数量(累计词频)	权重
1	性能质量	屏幕–性能–系统–电池	1 743	0.36
2	客服服务	服务–客服–态度	1 300	0.27
3	物流速度	快递–物流	578	0.12
4	诚实守信	正品–正版–原装	482	0.10
5	外形设计	外形–外观	424	0.09
6	产品价格	价–价格	289	0.06

建立的商家信誉维度体系如图 7 所示。从图 7 得到的信誉维度体系可以发现，1 性能质量、5 外形设计和 6 产品价格是关于产品自身的，2 客服服务、3 物流速度和 4 诚实守信是关于商家服务的。文献[6]也曾通过评论文本聚类研究数码产品信誉维度体系的建立，并将维度设置为 8 个，如图 8 所示。

对比图 7 和图 8 可以发现，图 7 中多了“外形设计”，而少了“交易安全性”、“售后服务”以及“品牌声誉”三项。分析原因笔者认为，关于“交易安全性”，随着网购交易形式的进化，尤其是第三方担保出现之后，

chinaXiv:201712.01380v1

ChinaXiv合作期刊

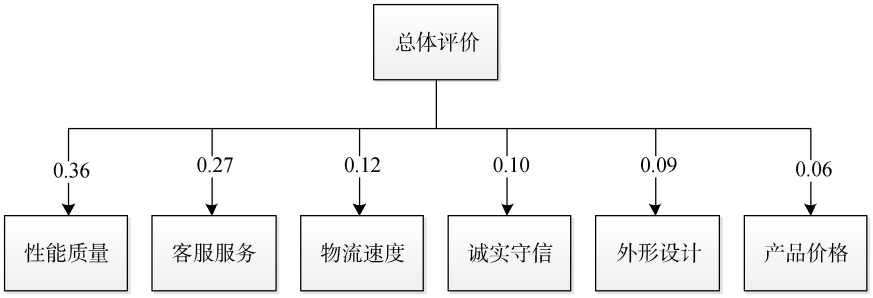

图 7　商家信誉维度体系

图 8　文献[6]构建的手机商家信誉维度

对于网购交易安全性的担忧已经越来越少，评论中已很少涉及到对于交易安全的担忧；关于“售后服务”，目前在大多数网购平台，手机这类由实体厂家生产的产品售后服务均由生产厂家提供，第三方网站作为中间媒介或担保的角色，因此该项在客户评论中也未体现；关于“品牌声誉”，文献[6]选取的评论文本来自于“中关村在线”网站，该网站相对于京东这种商务网站用户人群更为专业，评论更为深入，而且评论文本不仅包括手机，还包括电脑等其他数码产品，这些可能是其品牌声誉的来源，而从本文抽取的评论中无法体现出这一点；对于本文结果中多出的“外形设计”一项，说明随着手机出现时间的延续，各品牌在硬件性能方面逐渐趋同，用户消费能力不断提升，消费者在购买时不仅考量硬件方面，也正在越来越多关注外形设计等软实力方面。

7　结　语

在电子商务不断快速发展的背景下，商家数量的快速增长与商家水平的参差不齐，使得商家信誉评价问题越发紧迫与重要。用户评论文本是隐藏商家信誉信息的宝藏，随着 Web2.0 不断成熟，用户生成内容(UGC)呈现爆发式增长，热门商品下的评价数量成千上万，大量的评论文本在给用户带来重要信息的同时，也给用户的阅读浏览带来了很大的负担，使得用户无法快速获取有用信息。这也显示出评论文本挖掘工作的重要性和现实价值。

针对商家信誉评价的实际需要，以及评论文本区别于传统长文本的特点，本文主要在传统的基于词性词频的抽取方法基础上，对主题词进行扩展和词频调整，以发现隐藏的主题词信息；将 HNC 语义信息引入到改进的 CURE 聚类算法，将待聚类的主题词映射到 HNC 字词库，采用基于 HNC 符号的词语相似度计算，提出改进的 CURE 聚类算法；依据聚类簇中主题词的 HNC 符号提出类簇标签抽取方法，建立商家信誉维度，并计算出每个维度的权重；进行算法的实验验证和实例分析。

参考文献：

[1] 艾瑞咨询. 2016 年第三季度电子商务核心数据发布[EB/OL]. [2017-04-12]. http://report.iresearch.cn/content/2016/11/265616.shtml. (IResearch. The 3rd Quarter of 2016 E-commerce Core Data Release [EB/OL]. [2017-04-12]. http://report.iresearch.cn/content/2016/11/265616.shtml.)

[2] 中文互联网数据资讯中心. 2016 年全球互联网、社交媒体、移动设备普及情况报告[EB/OL]. [2017-04-12]. http://www.199it.com/ archives/437192.html. (Chinese Internet Data Center. 2016 Global Internet, Social Media, Mobile Device Popularity Report [EB/OL]. [2017-04-12]. http://www.199it.com/ archives/437192.html.)

[3] 鲁文. 社会化电子商务在线信誉的模型构建及实证研究[D]. 沈阳：沈阳工业大学，2015. (Lu Wen. Model Construction and Empirical Study on Online Reputation of Social Commerce [D]. Shenyang: Shenyang University of Technology, 2015.)

[4] 茹永梅. 基于模糊理论的 O2O 模式商家信誉评估模型[J]. 西安邮电大学学报，2016，21(3)：120-126. (Ru Yongmei. Reputation Evaluation Model for O2O Mode Businesses

chinaXiv:201712.01380v1

ChinaXiv合作期刊

Based on Fuzzy Theory[J]. Journal of Xi'an University of Posts and Telecommunications, 2016, 21(3): 120-126.)

[5] 吴维芳, 高宝俊, 杨海霞, 等. 评论文本对酒店满意度的影响: 基于情感分析的方法[J]. 数据分析与知识发现, 2017, 1(3): 62-71. (Wu Weifang, Gao Baojun, Yang Haixia, et al. The Impacts of Reviews on Hotel Satisfaction: A Sentiment Analysis Method [J]. Data Analysis and Knowledge Discovery, 2017, 1(3): 62-71.)

[6] 赵学锋, 陈传红, 陈获帆, 等. 基于文本聚类的电子零售商信誉维度发现研究[J]. 情报学报, 2011, 30(1): 63-75. (Zhao Xuefeng, Chen Chuanhong, Chen Huofan, et al. Study on the Discovery of Reputation Dimension of Online Merchants Based on Text-clustering [J]. Journal of the China Society for Scientific and Technical Information, 2011, 30(1): 63-75.)

[7] 赵学锋, 汤庆, 张睿, 等. 基于客户评论和语料库的在线酒店信誉维度挖掘[J]. 图书情报工作, 2012, 56(12): 124-129. (Zhao Xuefeng, Tang Qing, Zhang Rui, et al. Exploration of Dimensions of the Online Hotel Reputation Based on Customers' Text Comments and Corpus [J]. Library and Information Service, 2012, 56(12): 124-129.)

[8] 黄曾阳. HNC(概念层次网络)理论—计算机理解语言研究的新思路[M]. 北京: 清华大学出版社, 1998. (Huang Zengyang. HNC(Hierarchical Network of Concepts) —A New Approach to Computer Understanding Language Research [M]. Beijing: Tsinghua University Press, 1998.)

[9] Jiang M, Yan W, Wang X, et al. Wikipedia Based Approach for Clustering Keyword of Reviews[J]. Journal of Software, 2014, 9(9): 2246-2250.

[10] 耿焕同, 蔡庆生, 于琨, 等. 一种基于词共现图的文档主题词自动抽取方法[J]. 南京大学学报, 2006, 42(2): 156-162. (Geng Huantong, Cai Qingsheng, Yu Kun, et al. A Kind of Automatic Text Keyphrase Extraction Method Based on Word Co-occurrence [J]. Journal of Nanjing University, 2006, 42(2): 156-162.)

[11] 陈炯, 张永奎. 一种基于词聚类的中文文本主题抽取方法[J]. 计算机应用, 2005, 25(4): 754-756. (Chen Jiong, Zhang Yongkui. Novel Chinese Text Subject Extraction Method Based on Word Clustering[J]. Computer Applications, 2005, 25(4): 754-756.)

[12] 晋耀红. HNC(概念层次网络)语言理解技术及其应用[M]. 北京: 科学出版社, 2006. (Jin Yaohong. HNC (Hierarchical Network of Concepts) Language Understanding Technology and Its Applications[M]. Beijing: Science Press, 2006.)

[13] 苗传江. HNC(概念层次网络)理论导论[M]. 北京: 清华大学出版社, 2005. (Miao Chuanjiang. HNC (Hierarchical Network of Concepts) Introduction to the Theory [M]. Beijing: Tsinghua University Press, 2005.)

[14] 吴佐衍, 王宇. 基于 HNC 理论的词语相似度计算[J]. 中文信息学报, 2014, 28(2): 37-43. (Wu Zuoyan, Wang Yu. A New Measure of Semantic Similarity Based on Hierarchical Network of Concepts [J]. Journal of Chinese Information Processing, 2014, 28(2): 37-43.)

[15] 魏桂英, 郑玄轩. 层次聚类方法的 CURE 算法研究[J]. 科技和产业, 2005, 5(11): 24-26. (Wei Guiying, Zheng Xuanxuan. Research on CURE Algorithm of Hierarchical Clustering Method [J]. Science Technology and Industry, 2005, 5(11): 24-26.)

[16] 邵洪雨. 短文本聚类及聚类结果描述方法研究[D]. 大连: 大连理工大学, 2014. (Shao Hongyu. Research on Short Text Clustering and Cluster Description Method [D]. Dalian: Dalian University of Technology, 2014.)

[17] 夏天. 词向量聚类加权 TextRank 的关键词抽取[J]. 数据分析与知识发现, 2017, 1(2): 28-34. (Xia Tian. Extracting Keywords with Modified TextRank Model [J]. Data Analysis and Knowledge Discovery, 2017, 1(2): 28-34.)

[18] 陈毅恒. 文本检索结果聚类及类别标签抽取技术研究[D]. 哈尔滨: 哈尔滨工业大学, 2010. (Chen Yiheng. Research on Text Retrieval Results Clustering and Label Extraction[D]. Harbin: Harbin Institute of Technology, 2010.)

作者贡献声明:

王宇: 提出研究问题及研究思路, 设计研究方案;
李秀秀: 采集、分析数据, 起草、修订论文;
王宇, 李秀秀: 论文最终版本修订。

利益冲突声明:

所有作者声明不存在利益冲突关系。

支撑数据:

支撑数据由作者自存储, E-mail: 1434002120@qq.com。
[1] 李秀秀. TextReview.xlsx. 实例分析中抓取的原始产品评论数据集.
[2] 李秀秀. Review.xlsx. 实例分析中经过筛选处理的有效评论数据集.
[3] 李秀秀. KeyWords.xlsx. 实例分析中抽取的所有主题词.
[4] 李秀秀. ClusteringResults.xlsx. 实例分析中的主题词聚类结果.
[5] 李秀秀. HNCDatabase.mdb. HNC 符号映射中用到的 HNC 字词知识库.

收稿日期: 2017-05-27
收修改稿日期: 2017-07-23

chinaXiv:201712.01380v1

ChinaXiv合作期刊

Evaluating Business Reputation with E-Commerce Comments

Wang Yu　Li Xiuxiu
(Faculty of Management and Economics, Dalian University of Technology, Dalian 116024, China)

Abstract: [**Objective**] This paper proposes a new method to evaluate business reputation based on e-commerce comments. [**Methods**] First, we modified the key word extraction and clustering algorithm based on the HNC theory and text mining methods. Then, we extracted the cluster labels and calculated the weight of each cluster of the collected comments. [**Results**] We established a business reputation dimension system, with cellphone users' reviews posted on the Jingdong Online Shopping Platform. [**Limitations**] Some of the word symbols were generated manually due to the incomplete HNC thesaurus, which posed negative effects to larger-scale comments analysis. [**Conclusions**] The business reputation evaluation system can identify the commodity features that users really care about.

Keywords: Comment Texts　Topic Words Clustering　Reputation Dimension　E-Commerce

chinaXiv:201712.01380v1

Clarivate Analytics 与 Impactstory 合作支持科研人员将更便捷使用开放获取内容

Clarivate Analytics 于近日宣布与 Impactstory 开展全新的战略合作伙伴关系，这将为研究人员消除一道关键的障碍，即：高质量的、受信任的、经过同行评议的内容很少开放获取。根据双方合作伙伴关系，Clarivate Analytics 正在向 Impactstory 提供一项资助以建立 oaDOI 服务，从而使开放获取内容更容易被发现，研究工作从发现到发布变得更有效率。

科学出版是十分复杂的。在线搜索学术文章的研究人员很难获得可靠的有助于他们研究的搜索结果，这通常是因为搜索结果中省略了需要付费订阅的期刊文章，返回的是未经同行评议的版本或不违反版权法的版本。Clarivate Analytics 和 Impactstory 之间的合作关系将通过一种能在广泛的科学出版生态系统中持续发展的方法，为研究人员和各种机构提供对可信研究成果的开放获取。

oaDOI 服务来自非营利性组织 Impactstory。Impactstory 创建了一套在线工具，使得科学变得更加开放和可重用。目前，oaDOI 索引了 9 000 万篇文章，并通过一个免费、快速、开放的 API 提供开放获取的全文版本。Impactstory 还构建了 Unpaywall，这是一种免费的浏览器扩展，每当研究人员遇到付费文章时则使用 oaDOI 来查找全文。

Clarivate 正在开发和提供创新的分析和工作流程解决方案，从而提高整个研究生命周期的效率：从形成想法到实验验证，到同行评审、出版、传播和评估。与 Impactstory 的合作将研究人员连接到来自 Web of Science 的大约 1 800 万新的开放获取文章，从而加快 Clarivate 客户的创新发现阶段。此次合作尤其对于中小型机构将会特别有价值。

(编译自：http://news.clarivate.com/2017-06-23-Clarivate-Analytics-announces-landmark-partnership-with-Impactstory-to-make-open-access-content-easier-for-researchers-to-use?asPDF=1)

(本刊讯)